论文

论语义信息

石光明[1,2,3*], 高大化[2,3]

1. 鹏城实验室, 深圳 518055
2. 西安电子科技大学人工智能学院, 西安 710071
3. 琶洲实验室（黄埔）, 广州 510555
* 通信作者. E-mail: gmshi@xidian.edu.cn

国家自然科学基金 (批准号: 62293483, 62476206, 62101398), 国家重点研发计划 (批准号: 2022YFB2902900), 以及琶洲实验室 (黄埔)(批准号: 2022K0904) 资助项目

摘要 香农早在上世纪中提出了语法、语义、语用三个层面表达信息. 但由于当时对语义没有找到很好的数学刻画方法, 导致目前信息技术一直在语法层面表示信息. 信息技术停留在信号的感知、传输和处理层面, 缺乏对信号内容直接获取、传输和处理的理论方法. 从信号到内容始终存在语义鸿沟. 信号可以用数学函数表示, 形成了以香农信息论、奈奎斯特采样定理、傅里叶变换方法三座基石的信号处理理论. 而信号语义和内容至今还没有完好的数学表示, 导致难以跨越语义鸿沟, 更谈不上语义信息处理. 如何跨越这个鸿沟, 一直是计算、信息和智能领域共同研究的课题. 当今社会进入智能时代, 人机共处场景已近来临. 特别是当前随着语义通信的概念和技术方法的兴起, 如何让智能机器理解好信号内容是智能科技中的关键. 很多工科高校、科研院所的学者和大企业开发者对语义通信产生了浓厚的兴趣. 但从专业角度出发, 当前有关语义信息的概念非常不清晰, 没有建立统一公认的语义信息定义和刻画, 甚至有错误的观点, 更没有对信号内容的数学刻画. 本文讨论了信息的内涵, 对语义信息的基本概念, 语义信息物理产生过程、语义信息刻画和度量, 基于语义的信号信息表示、压缩、以及信号内容的数学刻画等给出了清晰的定义和明确计算方法. 希望形成语义信息处理理论, 深化和夯实智能通信和 AI 技术的理论基础.

关键词 语义信息, 语义通信, 语义刻画, 语义压缩, 信号内容

1 引言和动机

信息、物质、能源是组成我们世界的三大要素 [1]. 早在远古时代, 人们就认识到信息的作用. 为了能够成功狩猎, 古人通过观察动物留下的行动痕迹从中发现动物的位置和状态, 从而利于捕获猎物. 农耕时代人们观天象以确定播种和养护农作物时段, 以便丰收. 工业时代人们利用信息设计和制造各种产品, 满足人们现代生活需求. 那些时代的人们对信息有主观认知, 但没有从科学上给出信息的刻画, 人们发现和利用信息的能力处于较低的层级. 到了上世纪中叶, 著名的数学和信息学家香农

引用格式: 石光明, 高大化. 论语义信息.
Guangming SHI, Dahua GAO. On Semantic Information.

博士借用热力学中的熵概念给出了关于信息熵的物理和数学表述 [2], 即信息熵是事物状态不确定性或混乱度的刻画, 并用事物状态发生的概率来度量, 给出了度量信息熵 (也被称为信息量) 的计算公式. 从此, 人类社会进入信息时代, 人们传播信息、发现信息和利用信息的能力陡然增强. 特别是当今人类社会进入了智能时代, 对感知和认知以及使用信息, 到了更加深化的层面.

信息对社会发展的重要性是不言而喻的. 人类使用信息的目的是为了适应环境 (即改变我们自身、自我进化) 或改造环境 (亦称社会发展) 使之适应人类, 或者是满足人类认知世界以及情感交流. 一般而言, 这个过程可以分为四个步骤, 即感知、认知、决策、执行. 这四步是人类和智能体的智能行为的关键组成部分, 它们构成了一个连续的过程, 即 "感知—认知—决策—行动" 闭环.

" 感知" 是通过感官或传感器接收外界事物的信号或数据. 信息技术主要关注信号或数据层面. " 认知" 使智能体理解信息的含义, 形成知识和经验, 并做出合理决策. 然而, 目前缺乏有效的数学刻画, 因此这些问题尚未解决. 香农的信息熵度量信息量, 但未定义信号的内容、内涵、语义等. 没有这些基础问题的数学探讨, 机器无法有效认识世界. " 决策" 是在充分认知信息后, 结合任务目标和信息认知, 权衡利弊并做出行动策略. " 执行" 是将决策付诸实践, 采取行动. 随着时代发展, 通过通信技术, 人们可以远程感知信息. 所以上述的四个阶段可以扩充为 "感知—传输—认知—决策—行动" 五个阶段.

上述五个阶段, 核心问题是对信息内容的理解, 也就是对信号中的语义理解. **所谓内容就是智能体能够用先验知识对感知的特征信号进行解释和理解的结果, 在本文中将深入探讨.** 因为 "**信号保真保证信息保真**" 是信息时代的基本遵循, 所以信息时代的信息技术主要关注如何高保真传播、存储和展现信号. 对信号内容的解读都交给了人类大脑. 此时的通信技术追求的目标是如何把信源信号不失真且快速地传递到信宿端, 信宿端则要逼真再现信源端相同的信号; 而探测技术追求的也是不失真地感知事物状态信号. 不关心内容的信号处理技术不依靠先验信息, 导致的结果就是需要花费大资源代价 (如带宽、功耗) 等, 才能保证宽带信号保真和信息保真目的.

AI 科技诞生初期, 人们也试图让机器理解信号中的内容, 从最初的模式识别到神经网络, 再到深度学习的弱人工智能技术, 以及当今流行似乎具备了理解能力的基于大模型的 "强人工智能" 技术. 由于它们的理论基础是数学概率统计, 技术基础是聚类, 并没有涉及对信号或数据内容的直接理解和处理, 也就是说至今的 AI 并没有让机器真正理解信号或数据中的内容. 同样, 人类也无法真正理解 AI 是如何完成信息处理的, 导致 AI 的可解释性问题迄今为止依然悬而未决. 其实, AI 处理信息有其自身之道, 只是我们人类尚未明确. 其原因在于人与机器之间没有建立统一的信息语义表达.

要让机器理解信号 (或数据) 中的内容, 首先需要对信号或数据中的语义进行数学刻画, 然后才有可能用数学刻画信号中的内容, 也就是要回答信号的内容如何表示. 特别是在当前, 语义通信 [3∼14] 已成为通信和 AI 领域的热点问题. 语义通信是面向语义信息通信的方法, **但很多研究者对语义信息的理解有偏差**, 造成对某些语义通信和语义处理方法的不准确或不正确. 例如, 将一般的特征学习结果认定为语义特征, 导致的结果是人脑还是不能与机脑直接面向内容交流. 所以, 本文很有必要对语义信息进行论述.

2 语义信息的数学刻画

要论述语义信息的数学刻画, 首先需探讨语义的来源. 语义源于人脑对信号中信息的提取, 属于信息范畴, 为人脑服务. 人类借助信息掌握宇宙、自然和生命的规律, 理解并适应世界, 信息越清晰准确, 认知就越精确. 随着智能技术的发展, " 人脑" 一词可扩展为高智商智能体.

第二, 语义从何产生? 这个问题, 无论是信息学者还是脑神经科学家还没有给出明确的解释. 在此作者通过总结脑神经科学家们对语义研究工作[15∼21] 和对信息的认知, 提出了一种可理解的解释.

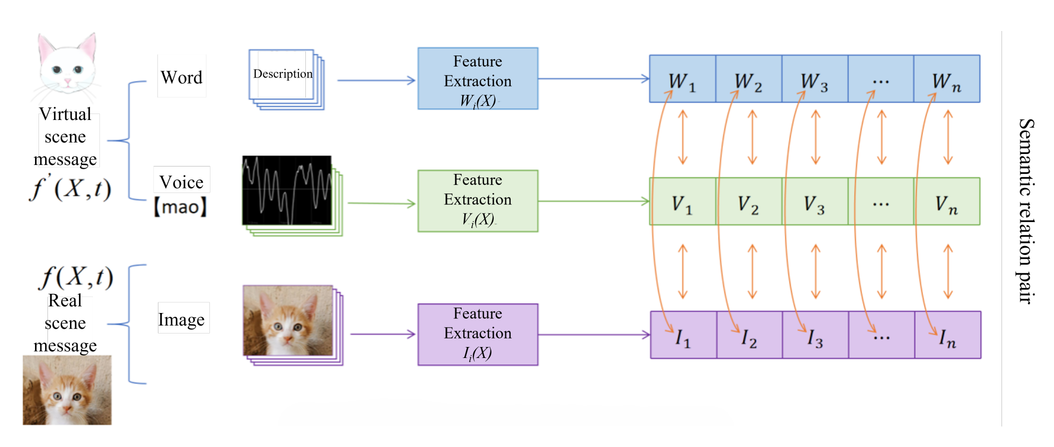


图 1 (网络版彩图) 语义：物景与像景信号相互标注

Figure 1 (Color online) Semantic: Mutual labeling between physical scenes and image signals.

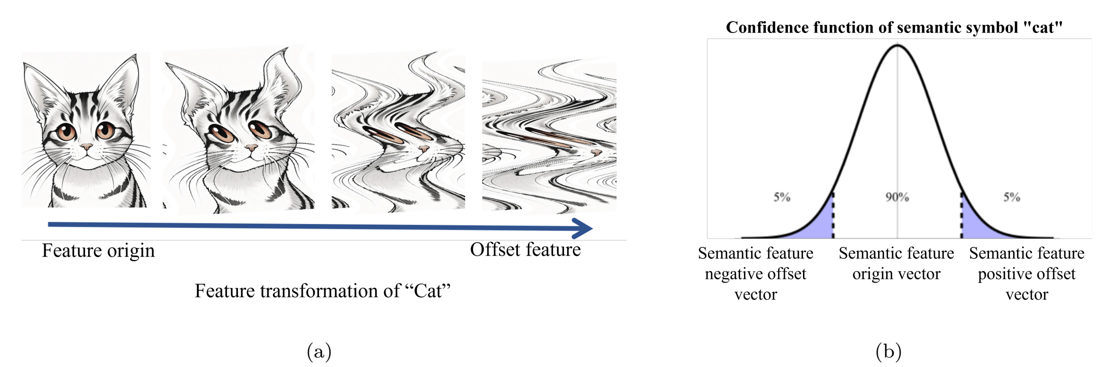
人脑是通过感觉器官感知信号, 语义是对信号含义的明确表达. 例如，对“火”的感觉来源于视觉、触觉和嗅觉，感知信号刺激大脑产生脉冲放电，形成语义。语义是基于感觉器官的高级感知，没有感觉就没有语义。例如，色弱者无法理解“红彤彤”的语义，盲人无法理解“白天”的完整含义，只能感知温暖。人类无法直接感觉电磁波，但通过其他物质变化和实验演示，我们能理解电磁波的语义。因此, 感觉器官是语义的基础, 有什么样的感觉器官就有相应模态的语义, 它们反应了事物状态的不同侧面.

第三, 语义在脑中如何表达? 脑神经科学家研究表明 [15,16], 概念在大脑中存在的基本单元是功能柱也称为皮层柱. 诺贝尔医学奖获奖者美国科学家 Hubel 和 Wiesel 博士 [15] 发现, 大脑视觉皮层中存在相同图像特征选择性和相同感受野位置的众多神经细胞, 以垂直于大脑表面的方式排列成柱状结构, 称为神经元功能柱 (Function of the column). 同一个功能柱内所有的神经细胞都编码相同的视觉信息, 它们只对某一种视觉特征发生反应, 从而形成该种视觉特征的基本单位. 他们的研究表明语义是人脑对客观事物的反映. 最新脑科学观点: 语义具有普适性, 而且是可表征的. 总之, 语义是大脑对客观事物状态的信号在脑中认知的反映 [16].

人的感觉器官收到外界事物状态信号, 引发产生的神经冲动, 通过神经元 (神经细胞) 组成的神经通路传导至大脑. 这些神经元通过轴突 (将神经冲动从细胞体传出) 和树突 (接收其他神经元传来的神经冲动) 相互连接. 经过初级感觉皮层处理进行初步的特征提取从初级感觉皮层提取的信息会进一步传递到联合皮层, 再联合皮层的整合与认知处理, 用于更高级的认知功能. 经过大脑多个区域的处理和整合后, 形成我们对外部世界的感知和意识. 最后将这些认知用脑控信号如语言、文字、脑电、图画等多种模态信号表达出来.

本文以脑为中心, 将信号分为两类: 一类是来自不受大脑控制的大自然事物状态信号, 称为物景信号. 如相机拍摄的图像、雷达探测到的电磁波、X 光图像、温度变化曲线等; 另一类是受大脑控制的像景信号，如语音、文字、手势、图画、脑电波和数学符号等. 语义是大脑对外界自然信号的反映，可以比喻为透镜成像过程，物景信号是”物”, 大脑是透镜, 像景信号是”像”. 这一过程是标注过程, 即用像景特征信号标注物景特征信号, 且通过共识形成语义. 在本文中, 特征信号是信号中的某些分量, 特征函数是特征库中的函数, 两者的数学表示相同. 语义是信号中的含义, 当前技术无法直接观测或度量, 可能未来通过脑电波测量. 本文所说的语义是各种信号中内容的含义, 它在大脑中有神经元对应表征, 是客观的 [16]. 但语义通过特征信号反映了事物的状态, 而特征信号是可观测的. 语义符号对应语义特征信号, 特征信号反映事物状态. 所以从特征信号的角度来看, 语义又是可以被观测的.

定义 1 语义是众人共识的用像景特征信号对物景特征信号的标注.



(a) (b)

图 2 (网络版彩图) 语义特征信号变化与语义符号的置信度关系

Figure 2 (Color online) The relationship between the variation of semantic feature signal and the confidence level of semantic symbols.

这些特征信号称为语义特征信号，所表达的信息即为语义信息。例如，图 1中，左侧为相机拍摄的图片，其语义通过中文发音【mao】或文字“猫”标注。对懂中文的人群而言，看到猫照片、听到发音【mao】或看到文字“猫”，都能理解其含义，具有相同的语义。对于英语使用者，猫的照片识别结果是“Cat”或发音【[kæt]】。这展示了语义形成的基本原理。

语义的特点是：一是物与像特征信号相互标注；二是群体共识。因此，严格来说，语义是通过共识标注事物状态的特征信号，而未被像景信号标注的特征信号不一定是语义特征信号。语义特征信号由物和像特征信号对 $s_i(x,t)$ 表示，所以借用复数的实部 (物景特征信号) 和虚部 (像景特征信号) 表示法, 可以记为:

$$s_i(x,t) = (f_i(x,t) + j \cdot g_i(x,t)) \rightarrow [S_i] \tag{1}$$

$$s_i \rightarrow f_i(x,n) + j * g_i(y,k) \tag{2}$$

其中 $f_i(x,t)$ 是物景信号中的特征信号, $g_i(x,t)$ 是像景信号中的特征信号, 公式 2是 **时** (n 表示离散时间点) **空** (k 表示离散空间点) 离散化后的表达. 为了便于记忆, 通常对所有语义特征信号用对应的语义符号集合 $[S_i]$ 标记语义特征信号. 公式 1或 2对物景语义特征信号、像景语义特征信号以及语义符号之间建立了数学联系. 所以, 所有包括中文文字在内的文字库就是一个文字语义符号集合, 同样所有的数学符号也是一个语义符号集合.

一个语义特征信号本质上刻画了事物的一种状态, 或者说事物一个状态可以用两种模态刻画或表述, 即连续的信号或数据和离散的符号. 一般而言, 物景信号是连续的、精确的信号, 传输、存储和记忆它的代价高; 而像景信号模态多样, 有连续信号如语音信号, 也有离散信号的如文字, 但它们表达的信息是离散的, 非精确的, 它们传输、存储和记忆代价低. 所以一个离散的像景符号与连续的物景特征信号不是一一对应的, 而是一个像景符号对应一类像景或物景信号. 例如, 从标准的猫头图像 (原位特征信号矢量) 到扭曲的猫头图像 (偏位特征信号矢量), 其特征信号发生变化, 那么“猫头”这个符号与这些特征信号矢量就存在置信度 (即概率) 分布:

$$C_{S_i(g_i(x,t))} = p(f_i(x,t)) \tag{3}$$

如图 2所示. 每一个语义符号与对应的语义特征信号的置信度 (Degree of Confidence) 曲线各不相同 (这主要与能识别的神经网络性能和特征信号的复杂程度相关), 但原位置信度高偏位置信度低的大致规律是相同的. 公式 1和 2表达的是语义符号与语义特征原位关系.

日常中, 我们常有这种感觉, 一个文字与另一个文字表示的含义类似或者几乎相同, 这是他们对应的特征信号的相似度决定的. 换句话说, 特征信号的相似度决定了语义符号含义的相似度, 这些语

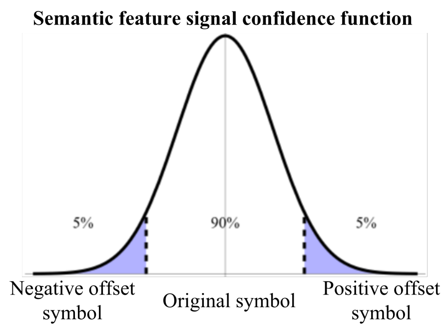


图 3 (网络版彩图) 语义符号变化与语义特征信号置信度关系
Figure 3 (Color online) The relationship between the variation of semantic symbols and the confidence level of semantic feature signals.

义符号也是同一类语义符号. 所以, 同样有另一个类似的置信度函数, 即一个特征信号对应一组 (同一类) 符号, 如图 3所示. 例如, 你好, 早安, 早, 嗨, 等都是问好, 打招呼的意思.

正如我们用离散的文字表示人类能够了解和理解的所有信息, 用离散语义符号表示信号中的信息也是一种高效表示信息的好方式, 或者说文字符号就是一种高效表达信息的方式. 语义符号本质是用离散的符号表示信号中的信息内容, 是从内容上对信号进行离散化. 这点与信号理论中基于奈奎斯特采样定理离散连续信号是不同的. 连续的物景信号按奈奎斯特采样定理离散化后, 还是描述事物的连续时空状态. 而用语义符号表示信息则是表达事物离散时空状态. 关于信号的语义离散化信息保真采样, 将在另一篇文章中给出方法, 即信号的**语义信息保真采样理论方法**.

3 消息、语义、信息、知识的关系

"消息、语义、信息和知识" 这四个词的应用非常广泛, 也非常普通. 在现实中, 各领域对信息一词的理解存在差异. 普通大众说的信息泛指消息之类. 在某些领域, 对信息又特指某类消息. 从哲学意义上来说, 信息是事物存在方式和运动规律的表现形式; 在通信领域中, 信息是用来消除不确定性的东西, 与事件 (状态) 发生的概率相关. 在探测领域, 信息是事件状态本身. 从认知科学角度看, 信息是主体所感知或所表述的事物存在的方式和运动状态. 而对"语义"的认知, 在百度词典中定义"语言所蕴含的意义就是语义". 语义学 (Semantics) 也作"语意学", 是一个涉及到语言学、逻辑学、计算机科学、自然语言处理、认知科学、心理学等诸多领域的术语. 这些定义或解释似乎能让大家明白, 但不具备信息工程上的操控性. 在语义通信领域中的语义是指各种多模态信号中内容的含义. **本文中的语义信息是指在各种多模态信号中能够被人脑感知出含义的消息**.

人脑在接收消息信号后首先对信号分割, 逐个解码识别成语义符号, 然后分析这些符号的关系, 最后对所有语义符号整体理解, 获得其含义. 所谓**含义, 就是消息所表示的事物状态在大脑中形成的像**, 是大脑对事物状态的表示, 犹如照镜子. 信息来源于消息, 是人脑对消息感知后的结果. **信息本质是消息接收者事先未知且能解码理解其含义的那部分消息**. 如果接收者在接收消息之前就已知该消息的所有内容, 则这次收到的消息对他而言就没有信息, 可以认为**已知的消息就是知识**. 所谓能够解码语义, 就是接收者能够识别相应的编码符号, 这需要接收者事先进行训练, 然后接收者再利用已有的知识库解码语义符号组的含义. **消息中还有接收者不能用已知符号和知识库解码的部分, 称为暗 (Dark) 消息, 因为没有先验信息帮助解码.** 例如，路人甲通过声音传递消息给路人乙，乙可能能听懂一部分已知内容，另一部分是未曾知道的，而有些内容乙则完全无法理解。这涉及三个层面：第一，信号层，乙能听到甲发出的声音；第二，语义层，乙能从信号中提取语义，如果乙不懂英文，甲用英文说话时乙无法理解；第三，信息层，乙能够解码每个字，但若无法理解字的组合内容，比如甲向小学生讲解相对论，乙无法获得信息，尽管每个字都懂。这些层次的区分反映了消息、信息、知

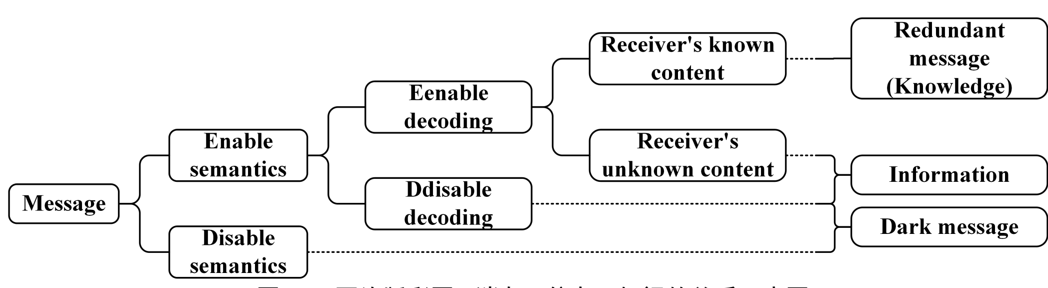


图 4 (网络版彩图) 消息、信息、知识的关系示意图

Figure 4 (Color online) A diagram illustrating the relationship between message, information, and knowledge.

识和暗消息之间的关系，见图 4。

所以, 知识对应消息中已知部分的内容; 语义是对应消息中能够被识别的部分内容; 信息是接收者事先未知且能理解含义的部分内容. 信息被接收者吸收转化为新知识用于后续通信. 信息是接收方接收消息前后的知识差异, 这个差异用于消除接收者对原未知事物认识的不确定性. 对于接收者而言, 一个消息中有没有信息取决于它是否能解码消息中的语义, 以及这些语义综合在一起所表示的含义事先是否已知. 由此得出结论:

在语义的范畴内, 对接收者而言, 一个消息包含了知识、信息和暗消息.

目前的通信技术是对发送方而言的信息, 常表达发送了多少信息. 实际上, 信息应该对接收方而言, 要表达成接收到了多少信息. 这个重要的观念正是通信本质所在. 通信的目的是对方能够收到多少信息而不是发送方发送了多少信息. 探测技术的本质是能够探测多少信息而不是能够解码多少信息. 接下来再讨论知识. 从图 4可知, 知识是过去的信息.

定理 1 一个新知识都有由旧知识构成的.

假设, 新知识用 ${k_i}^{(1)}$, 是由 n 个旧知识 ${k_j}^{(0)}$ 结构而成, 用公式 4表示:

$$ {k_i}^1 = \bigcup_{j=1}^{n} {k_j}^{(0)} \tag{4} $$

其中, $\bigcup\limits_{j=1}^{n}$ 算符表示某种结构组合运算. 公式 4证明成立如下:

(1) 因为知识都是能够被人理解的, 这是知识的前提条件.

(2) 因此, 假设新知识不全是旧知识某种结构组织而成, 其中包含了未知内容的或者还没有确认的新知识本身;

(3) 则该新知识就不能被人理解, 也就不能称为知识.

(4) 推理结论与知识的前提矛盾.

(5) 因此, 公式 4正确.

可知, 新知识是由旧知识编码而成, 科研工作者的工作内容之一是挖掘或构造新知识的编码方式. 古希腊的哲学家和科学家相信[23], 世界上始终存在一个必然正确的元起点, 从这个元起点出发, 通过逻辑性的推导, 人们就可以获得新知识. 公式 4正好与此观点相同.

本文没有给出旧知识是如何编码构造新知识, 这将在另一篇论文中讨论. 从信息的角度看人类的历史, 就是不断用过去的先验信息辅助解译当下信号、获得信息、形成知识, 再推动社会前进的发展史, 循环往复. 先验信息 (也称为知识) 非常重要, 面向信号的信息技术很少利用先验或者没有充分利用先验, 而面向内容的信息技术, 即基于语义信息技术就是要充分利用先验技术. 总之, 能够解码信号中内容的前提是有语义特征信号的先验. 提取信号内容的过程从数学上讲, 实质上是条件概率问题.

4 信号内容的数学刻画

一个信号反映了接收者可理解的事物状态就有了内容. 内容是反映事物状态的消息. 所谓信号内容的表示是通过语义符号或语义特征信号刻画信号内容. 传统信号中的内容是由人脑利用自己的先验知识从中提取, 再转成语义符号表示. 而至今机器是不理解信号中的内容, 因为对内容没有给出可计算操作的方法. 当前, 几类方法与内容表示相关.

一是**基于概率统计的方法**. 二是**信息论方法**. 三是**特征提取方法**. 四是**深度学习方法**.

目前的信号内容刻画方法尚不完美，且与人脑感知的信号内容不同。它们仅提取信号的某些分布或特征，未直接反映事物状态。人脑表达信号内容有两种方式：一种是通过语义特征函数的实部或虚部，另一种是通过语义符号。本文提出模仿人脑的两种刻画方式：用语义特征信号或语义符号表示信号内容，首先研究语义特征信号的表示方法。

一般而言, 对信号中内容的探究, 有两个视角, 一是从信号自身角度, 即信号拥有多少内容; 二是从接收信号角度, 即接收者能够读到多少内容. 现实中信号内容只有被接收者感知到才有意义, 因此本文主要从第二个角度来开展研究. 而第一角度是则包含了对未知内容的探究. 众所周知, 信号承载了消息, 感知到信号就是收到一个消息, 特征信号表示消息内容. 有的特征信号不能被理解, 称为暗特征信号; 有的特征信号能够被人理解, 称为语义特征信号. **由信号中所有语义特征信号及其相互关系构成的含义称为信号内容**. 接收者收到一个消息且能够消除他对某个事件认知上的不确定性, 则这些消息中有信息. 如果没有消除接收者在认知上的不确定性, 则该消息就只是消息.

根据香农信息熵理论, 信息是与事件状态发生概率密切相关, 它是信源 (事件) 自身的状态发生的可能性体现, 与接收者无关. 这个信息量对接收者而言实质是消息量. 例如, 假设两个独立事件 A 和 B, 事件 A 发生状态的 a1、a2、a3 概率与事件 B 发生状态的 b1、b2、b3 概率都为 1/3, 对通信而言它们的信息量是一样的, 但对接收者而言, 事件 A 与事件 B 是两个不同的事件. 我们应该关注事件状态发生的概率, 还是应该关注状态自身, 即 A 或 B 是具体的状态 (消息的内容)? 通信技术关注的是能传递多少信息量或消息量, A 和 B 的状态由接收者来解读或判断; 而探测技术关注的是探测到的事物是什么状态, 不管接收者是否能够判断出 A 或 B, 探测技术是希望能够完全保真传递探测到的信号. 例如, 运送 1Kg 水和 1Kg 铁, 它们重量一样 (类似信息熵一样), 但物质属性不同 (类似事物状态不同), 运输车关心的是重量数, 而使用者关心的是物质属性和重量. 通信技术关注的是传多少, 而探测技术关心的是探测到了什么. 而为智能服务的通信技术最终目标还是要给接收方传送它能懂得的有用信息, 或者说接收者看重的是接收到的内容. 所以通信最终目的应该是关心传什么和传多少两个问题. 传多少的问题, 由香农信息熵理论给予指导解决, 但传什么的问题还没有理论指导, 因为信号内容还没有数学刻画方法. 本文将回答此问题. 什么是信号内容?

定义 2 信号的内容是指其刻画事物状态的所有特征信号.

4.1 物景信号内容的数学刻画

假设一个多维时空物景信号 $y(x,t)$,((x,t) 为空时维度), 可以分解为可识别的 n 特征信号 $f_i(x,t)$ 的变化 (即平移、旋转和缩放等) 信号, k 个不能识别的特征信号 $D_j(x,t)$(后文将讨论对此处理方法), 以及不表示任何消息的信号 $e(x,t)$, 即:

$$y(x,t)=\sum_{i=1}^{n}\beta_i\cdot\rho_i(a_if_i(x-x_i,t-t_i),\theta_i)+\sum_{j=1}^{k}D_j(x,t)+e(x,t) \tag{5}$$

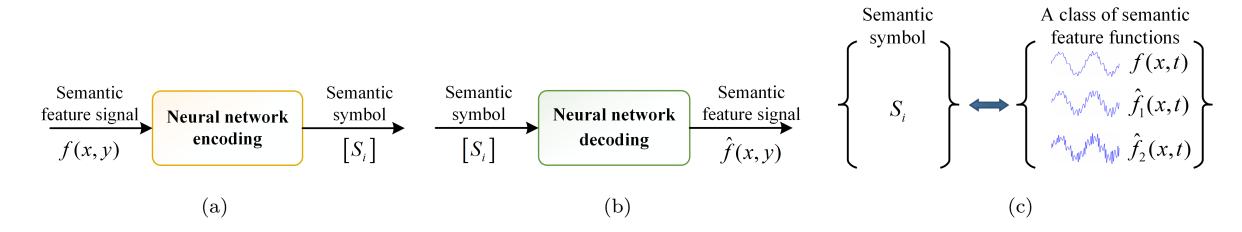


图 5 (网络版彩图) 语义特征与语义符号一一对应关系

Figure 5 (Color online) The one-to-one correspondence between semantic features and semantic symbols.

其中, α_i, β_i 是特征信号缩放系数, ρ_i, θ_i 分别是旋转变化函数和旋转角度, x_i, t_i 是特征函数的时空位置, 这些被称为特征函数的属性参数. 这些属性参数不同会引起语义特征信号变化. 在此, 本文给出语义的原位和偏位信号定义.

定义 3 语义原位特征信号是标准的语义特征信号. 由于原位特征信号属性参数变化引起变化的特征信号, 且可被识别为原位特征信号, 称为偏位语义特征信号.

定理 2 信号内容是由语义特征原位信号及其偏位信号编码而成.

证明过程如下:

(1) 众所周知, 所谓内容就是能够接收者识别含义的消息.

(2) 如果信号是用语义特征信号其偏位信号的时空关系编码, 则能够被接收者识别含义;

(3) 如果信号是用非语义特征信号和不能识别的其他特征信号编码, 则不能够被接收者识别含义, 也就不能称为内容;

(4) 所以定理 2成立.

公式 5表达了信号用语义特征编码的方法, 原位或偏位语义特征信号是事先已知的, 而特征函数的属性参数是事先未知的, 这点正体现信息的随机性. 这种由可识别的特征函数的属性参数随机变化编码的信号, 其原理与用 0、1 比特随机排位编码信号一样, 都是用已知的符号或函数, 通过符号随机顺序或参数变化表达信息. 可识别的特征函数已知, 如同 0、1 符号是已知的一样.

人脑感知到 $f_i(x,t)$ 后就有语义, 多个 $\sum_{i=1}^{n}\beta_i\cdot\rho_i(a_if_i(x-x_i,t-t_i),\theta_i)$ 理解后会形成含义, 含义是由多个语义单元组成的整体内容. 为什么人脑感受到语义特征信号后就会有语义感觉? 这是因为人脑利用了语义特征先验知识, 事先大脑经过了训练, 对该语义特征信号与某个事物的状态进行了关联. 例如, 父母对小孩大脑最初训练是从 "什么是什么" 开始的, 即手指向某物体, 口中发出该物体的名称声音. 这些训练结果就是小孩的基本知识, 也是他长大过程中其脑学习的基础或者先验知识.

如果信号中有能懂的内容, 那么这个内容与先验知识必然相关, 也就是说信号内容与先验知识不是独立分布的, 信号内容是先验的条件概率.

借用信息熵的概念, 刻画信号内容熵为随机变量, 假设 $y(x,t)$, 其中 t,x 代表时空维度, 在给定随机变量 $f_i(x,t)$ 的条件下, 随机变量 $y(x,t)$ 的条件熵, 定义为:

$$H(\mathrm{Y}(x,t)/\mathrm{f}(x,t))=\sum_{f\in F}p(f(x,t))H(\mathrm{Y}(x,t)|\mathrm{f}(x,t)=\mathrm{f}_i(x,t)) \tag{6}$$

信号内容熵仍然是消息量的概念, 根据条件熵公式推导, 可知:

$$H(\mathrm{Y}(x,t)/\mathrm{f}(x,t))\leqslant H(\mathrm{Y}(x,t)) \tag{7}$$

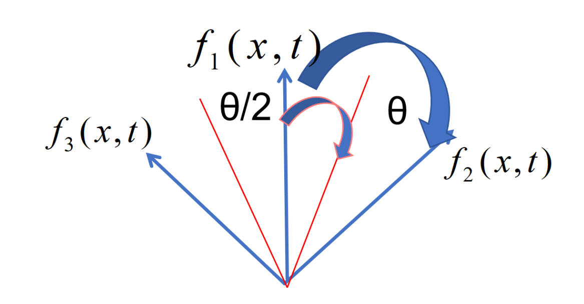


图 6 (网络版彩图) 语义特征函数原位与偏位夹角关系

Figure 6 (Color online) The in-place and offset angular relationship of the semantic feature function.

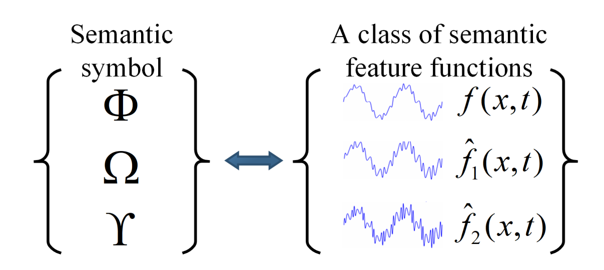


图 7 (网络版彩图) 语义特征函数原位与偏位夹角关系

Figure 7 (Color online) The many-to-one relationship between semantic features and semantic symbols.

即内容熵小于等于消息熵. 由上文可知, 内容具备已知的原位语义或偏位语义特征函数. 特别地, 针对信息不确定特征, 即信号内容如何消除接收者不确定性? 其实, 在公式 5已经表明, 其中的参数 α_i, β_i、ρ_i, θ_i、x_i, t_i 的随机性表现为信息的不确定性特征.

以上讨论了信号的语义编码方式. 实际中, 常需从信号 $y(x,t)$ 中确定与内容相关的语义特征信号. 本文提出通过最小化 $y(x,t)$ 与所有可能的语义特征信号之差平方进行求解, 表达式为:

$$\min \left| y(x,t) - \sum_{i=1}^{n} \beta \cdot \rho(\alpha f_i(X-x, T-t), \theta) \right|^2 \tag{8}$$

公式 8仅在剔除中所有语义特征信号原位或者偏位后才能获得最小值. 换句话说, 通过求最小值就可以找到信号所有表达内容的特征信号.

下面讨论对信号中包含了非语义特征信号的内容表达方法. 对不能识别的特征信号 $D_j(x,t)$, 可以采取进一步分解的模式处理, 即将特征信号 $D_j(x,t)$ 向下沿已知的语义子特征函数方向分解, 直至分解到语义子特征函数, 然后采用这些子特征函数的集合定义 $D_j(x,t)$, 并定义一个对应的符号名, 即其像信号.

$$D_{\mathrm{j}}(x,n) = \sum_{p=1}^{m} \alpha_p \times f'_{\mathrm{j},p}(x,n) = \sum_{p=1}^{m} \alpha_p \sum_{q=1}^{l} \beta_{p,q} f''_{\ \mathrm{j},p,q}(x,n) = ...(9)$$

其中, $f'_{i,p}(x,n)$, $f''_{\ \mathrm{j},p,q}(x,n)$, 分别为子特征信号、子子特征信号, α, β 为系数. 下面举例说明公式 9的含义. 例如, 有一种动物称为麋鹿, 最初人们见到它时, 不知到它该称为什么, 对人而言就是 $D_j(x,t)$. 因为它头脸狭长像马、蹄子宽大像牛、尾细长像驴、角像鹿又与其它鹿略有不同. 人们通过分解它各子部分, 分别与人类认知中的马、牛、驴、鹿的各自子部分相近似 (子部分特征信号的原位或偏位), 对应已知的子特征信号 $f'_{i,p}(x,n)$ 或 $f''_{\ \mathrm{j},p,q}(x,n)$ 综合在一起, 因此又名四不像, 并约定了一个新的符号 "麋鹿" 标注这个新的特征.

所以, 对于 $D_j(x,t)$ 进行特征分解的分解, 直到分解的结果与已知的子特征函数库或子子特征函数相似. 从图像信号的角度看, 组成信号的最小基元是像素. 这种特征分解到终点就是像素. 所以组成 $D_j(x,t)$ 的子子特征信号总是存在的, 关键是这些子子特征信号在哪个层级. 最低层级就是像素级.

在此说明, 拥有的语义知识库中的每一个语义特征函数 $f_i(x,t)$, 可能是目标级的特征, 也可能是部件级或子部件级特征函数 $f_i^{'(x,t)}$, 被称为语义基元函数. 也就是说, 一个未知的特征函数 $D_j(x,t)$, 可以近似分解为语义基元加权和.

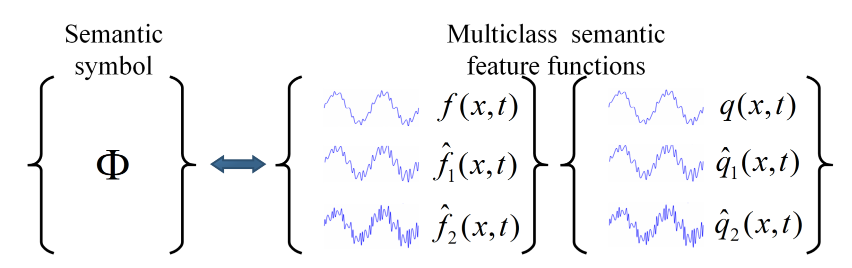


图 8 (网络版彩图) 语义符号对应多个语义特征函数

Figure 8 (Color online) A semantic symbol corresponds to multiple semantic feature functions.s

4.2 像景信号内容的数学刻画

像景信号有两类, 一类是连续信号, 如语音、脑电等; 另一类是离散信号, 如文字和包括数学符号在内的各种符号. 无论是连续信号还是离散符号, 它们实际表达信息都是离散的, 也就是说语音以连续信号表达了事物离散状态, 即语音与离散文字是等效的. 像景信号以离散的方式表达事物状态正是大脑明智所在. 大脑不需要存储和记忆大量原始事物状态的细节信息, 这样可以低代价低能耗存储和记忆事物状态.

由离散符号构成的信号, 可以说其每一个符号都是信号的内容, 因为这些符号是人脑对物景信号加工后的结果. 用这些符号的先后顺序变化编码可以传达出信息内涵. 语言就是用文字符号编码表达信息的一种方式. 著名哲学家维特根斯坦说过“语言的界限即是世界的界限”. 换句话说, 语言可以描述世界的所有事物时空状态, 语言和文字与物景信号有对应关系. 大众已知, 一句话的标准格式是主谓宾定补状, 它们分别是名词、动词、形容词、副词和数字等符号. 用这种结构将这些属性符号组合在一起形成表达事物状态的语言编码. 主语 (subject, $S(x,t)$) 和宾语 (object, $O(x,t)$) 在动词 (verb) 谓语表达了发生了相应的相互作用. 主语和宾语前面的形容词是系数 α,β, 它们作用的程度是用副词 δ 进行表达, 状语和补语则代表函数的取值域. 这种编码能够重构出事物状态 $y(x,t)$, 有以下对应关系:

$$y(x,t) = \delta \cdot V(\alpha \cdot S(x,t), \beta \cdot O(x,t))|_{\lambda}^{\gamma} \tag{10}$$

公式 10可以解决用自然语言描述的知识转化成物景信号, 从而可以将知识转化成数据, 弥补 AI 训练中缺少有效数据问题.

5 语义知识库的结构

在表示信号内容时, 从约定好的最基本的单个符号到几个符号组合的词, 再到一段符号组成的段落, 其含义都依赖已建立的知识库的内容. 符号只是一个代表, 具体代表什么, 则依赖知识库. 知识库就是一个翻译的码本. 但需要注意的是, 这个码本是直接面向内容, 而且是可动态扩充的.

知识描述了人类已知或可知的主客观事物状态的时空运动规律. 知识具有层级化特点, 从最简单最浅显的知识节点开始, 逐步累叠形成高层知识, 或称为高深知识. 最基本的知识可以称为基元知识. 表达一个知识节点, 有如下几个要素:

(1) 目标对象, 可能涉及多个对象;

(2) 对象的属性, 包括时、空坐标位置、旋转方向、尺寸大小、缩放比例、颜色、物态等;

(3) 对象之间的行为关系, 或者随时间变化的目标对象的空间位置和形态等;

(4) 变化的属性, 变化的速率、变化的方位等.

这一个节点可以用一个新符号命名, 又称为一个新目标. 它和其它目标相互作用关系又会再得到一个新目标. 用层级化方式循环迭代, 周而复始, 从而形成更高层的符号. 公式4已经表达了这个概

$y(x,t)$ source → Recognition filtering NN → Encode into semantic symbols and symbol relations $[S_i]$ → Refined down sample NN → High-level semantic symbols

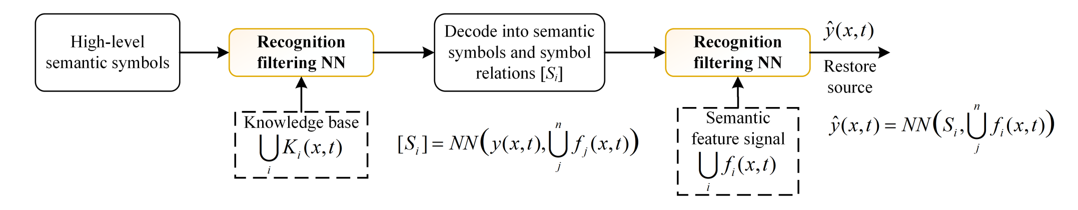
Semantic feature signal $\bigcup_i f_i(x,t)$

$[S_i] = NN\left(y(x,t), \bigcup_j^n f_j(x,t)\right)$

Knowledge base $\bigcup_i K_i(x,t)$

$[S_j] = NN\left(S_i, \bigcup_j^n K_i(x,t)\right)$

(a) 基于语义先验且信号内容保真的超高压缩率的信号压缩框图



(b) 基于语义先验且信号内容保真的超高压缩率的信号解压框图

图 9 (网络版彩图) 面向语义信息保真的信号压缩方法

Figure 9 (Color online) A signal compression method for semantic information fidelity.

念. 从这个意义上说, 知识是可以计算的, 也是可以通过计算自动生成的. 用计算方式发现或者创造知识, 也正是 AI 技术可以做的工作.

6 语义一致性的度量方法

大多数情况, 语义符号与语义特征信号是一一相互标注或称为一一对应的. 但现实中, 毕竟语义与大脑有关, 大脑对物体识别具有很好的泛化能力, 对存在一定偏差的语义特征信号也能够识别, 所以语义符号不是与一个语义特征信号对应, 而是与一类特征信号对应. 在第 2节中引入置信度函数, 从概念上刻画的是同一类语义特征信号或一类语义符号偏差的问题. 而在此讨论的是如何计算语义特征的偏差, 也就是语义一致性的度量.

语义符号与语义特征信号存在三种对应关系. 一是一个语义符号对应一类语义特征信号; 二是多个语义符号对应一类语义特征信号; 三是一个语义符号对应多类语义特征信号. 人脑是利用上下文场景将 "一对多" 或 "多对一" 明确成 "一对一" . 如果没有上下文, 就会出差错.

6.1 一对一类的语义特征函数偏差计算

首先研究语义符号 $[S_i]$ 与同一类语义特征信号对应的情况, 如图 5(c). 语义特征信号与语义符号是通过神经网络解码和编码的相互转化, 如图 5(a)和 5(b)所示, 一定程度偏差的语义特征信号也可以被正确解码. 能解码的偏差度大小取决于神经网络的泛化能力. 故定义标准的语义特征信号为原位特征信号 $f(x,t)$, 与原位存在偏差的语义特征信号称为偏位特征信号 $\widehat{f}(x,t)$.

在此, 分别选用两种方法计算原位与偏差误差. 一是均方误差 MSE $d(f(x,t), \widehat{f}(x,t))$, 这反映它们在时空域的差异性, 如公式 11.

$$d(f(x,t), \hat{f}(x,t)) = (f(x,t) - \hat{f}(x,t))^2 \tag{11}$$

二是计算矢量夹角 θ. 在同一类语义特征函数中, 它们不是正交的. 如果把它们看成一个矢量, 那么他们相互有一定的夹角 (0 180^0), 如公式 12,

表 1 水声语义通信图像传输测试与抗信道衰落能力结果

Table 1 Test results of underwater acoustic semantic communication image transmission and resistance to channel fading.

图像组序号	数量	所提语义方法		传统方法 (JPEG)		压缩比倍增数
		压缩比	语义保真率	压缩比	语义保真率	
1	15	1588.97	97.65 %	77.65	0.00%	20.46
2	15	1347.97	97.91 %	65.79	5.30 %	20.48
3	15	1430.07	98.41 %	69.18	15.40 %	20.66
4	15	1329.89	98.27 %	64.70	10.58 %	20.55
5	15	1509.52	98.04 %	74.36	15.59 %	20.29
6	15	480.03	97.76 %	23.35	5.93 %	20.55
7	15	636.27	98.36 %	31.38	20.52 %	20.27
8	15	477.38	97.73 %	23.37	0.00 %	20.42
9	15	574.91	98.30 %	28.16	5.45 %	20.41

$$\cos(\theta) = \frac{\vec{f}(x,t) \cdot \vec{\tilde{f}}(x,t)}{\left|\vec{f}(x,t)\right| \left|\vec{\tilde{f}}(x,t)\right|} \tag{12}$$

语义的一致性度量与 AI 的能力有密切相关. 如果 AI 分辨两个语义特征函数矢量的泛化能力很弱, 那么原位与偏位的特征函数偏差就不能过大, 反之亦反. 假设 AI 网络能够鉴别的相距最近的两个原位语义特征函数夹角为 θ, 那么偏位特征函数最大可偏的角度就为 $\theta/2$. 所以一致性的语义特征函数之间的矢量夹角偏差允许在 $\theta/2$ 以内. 也就是说, 如果偏位语义特征函数矢量与原位语义特征函数矢量在 $\theta/2$ 之内, 那么此偏位与原位的语义是一致的.

总之, 两个语义是否一致, 取决与所用 AI 网络的鉴别语义特征函数的偏差能力.

- 多个语义符号对应一类语义特征函数的偏差计算: 多个词语可以表示相同含义, 如,“你好”、“早安”、“早” 都有问好的含义, 如图 7所示. 它们的语义特征一致性通过公式 11和 12计算.

- 一个语义符号对应多类语义特征函数的偏差计算: 一个语义符号对应多类语义特征的情况也常见, 如 “行” 有 “同意”、“行走”、“行业” 等不同含义. 对此, 需要引入上下文信息, 基于上下文建立 AI 网络判断语义符号对应的语义特征信号, 如图 8所示. 确定类别后, 再用公式 11和 12计算语义特征的偏差.

7 语义保真率计算

上文主要是从语义特征函数方面讨论了语义一致性问题, 本文还从信息保真角度给出语义保真率度量方法. 一般, 信源信号经过压缩、传播、重构等过程会出现信号失真, 信息损失等. 传统用信号失真度或比特失真率来度量, 但不能度量出信息损失程度, 因为信号失真不直接反映信息失真. 为了度量信息失真度, 本文给出语义保真率的计算方法. 用 S_R 表示语义保真率, 用 S_C 表示信源内容, 用 S_O 表示信宿内容, 则语义保真率按公式 13计算:

$$S_R = \frac{S_O}{S_C} \tag{13}$$

正如前文所述, 信源的内容是用语义特征信号表达的. 所谓信息保真, 本意应该对接收者而言, 信宿内容占信源内容的比率. 这里涉及到内容量如何计算的问题. 以图像信源为例, 从内容上看, 组成图像的内容是其中可识别的目标语义特征信号及其时空关系. 而语义特征信号又可以转化成对应

的语言符号. 所以, 刻画图像的内容可以用语义符号表达. 符号数多, 表达的内容就多, 当然其中不能有冗余的符号, 也就是说一个语义特征信号用一个语义符号 (对应语言是一个词, 不是一个字); 一个语义特征与另一个语义特征时空关系用一组符号表示.

在自然语言处理 (NLP) 中, 常用双语评估基准 (Biingual Evaluation Understudy, BLUE) 方法判断两个文句表示的含义是否一致. BLEU 是一组度量机器翻译和自然语言生成型性能的评估指标, BLEU 通过计算 N-gram(连续 N 个词) 的匹配程度, 来评估机器翻译的精确率 (Precision), 侧重于衡量机器翻译输出与参考翻译之间的相似程度.

人脑对语义符号的一致性有很好的判断方法, 通过对符号的主观评分 (MOS) 判断语义是否一致. 受 BLEU 启发, 本文设计了一种语言特征信号翻译成语义符号, 并按 BLEU 原则判断语义符号精确匹配的方法, 称为“特征符号转化法”. 具体的计算算法如下:

(1) 设定需要分割的最小目标, 即观测信息的微观尺度 Scale;

(2) 在给定的 Scale, 对信源中目标逐个分割;

(3) 逐个目标特征信号识别转化成语义符号, 包括目标所在的时空位置, 也就对应公式 8中的内容描述参数;

(4) 用符号表达信源获得 S_C;

(5) 同样对信宿重复上述第 2 至 4 步获得 S_O;

(6) 按 BLEU n-gram 方式计算语义保真率.

8 基于语义信息的信号表示内容保真压缩方法

传统信号压缩方法是用时域、频域或小波域表示同一信号, 通过减少信号在这些域的冗余性, 从而达到信号压缩的目的. 它们都是采用面向信号保真原则表示信号, 即无论什么方法表示信号, 它们都要求不失真地相互转换成原始信号, 也就是说编码和解码要完全或近似完全重构. 那么, **能否做到压缩后尽管信号失真了但内容还保真？这样可以非常明显地提高信号压缩水平**. 回答是肯定的.

据上文论述, 信号承载内容, 内容可以用语义特征信号表示, 语义特征信号又可以转换为语义符号, 而语义符号又可以借助知识库为支撑, 进一步凝练, 用更少的符号来表达, 从而实现信号被大比例压缩. 其解压方法存在两种方式. 一是基于符号和知识库的重建; 二是基于符号和知识的生成. 这种信号压缩方式其原则是保证信号的语义信息或者内容保真.

消息的内容具有可以被表达或被代表的特征. 当然, 这个凝练的基础就是语义知识库, 因为知识库中包含了很多内容. 这也是为什么用像景语义特征或者语义符号可以大比例压缩信号的原因.

本文在此提出了面向语义信息保真的信号超高倍数的信号压缩方法, 其具体的算法如下:

■ 编码算法

(1) 按目标分割信源;

(2) 基于深度网络识别信源;

(3) 将识别结果转用符号表达;

(4) 分析识别后的所有符号, 建立它们之间的时空结构关系;

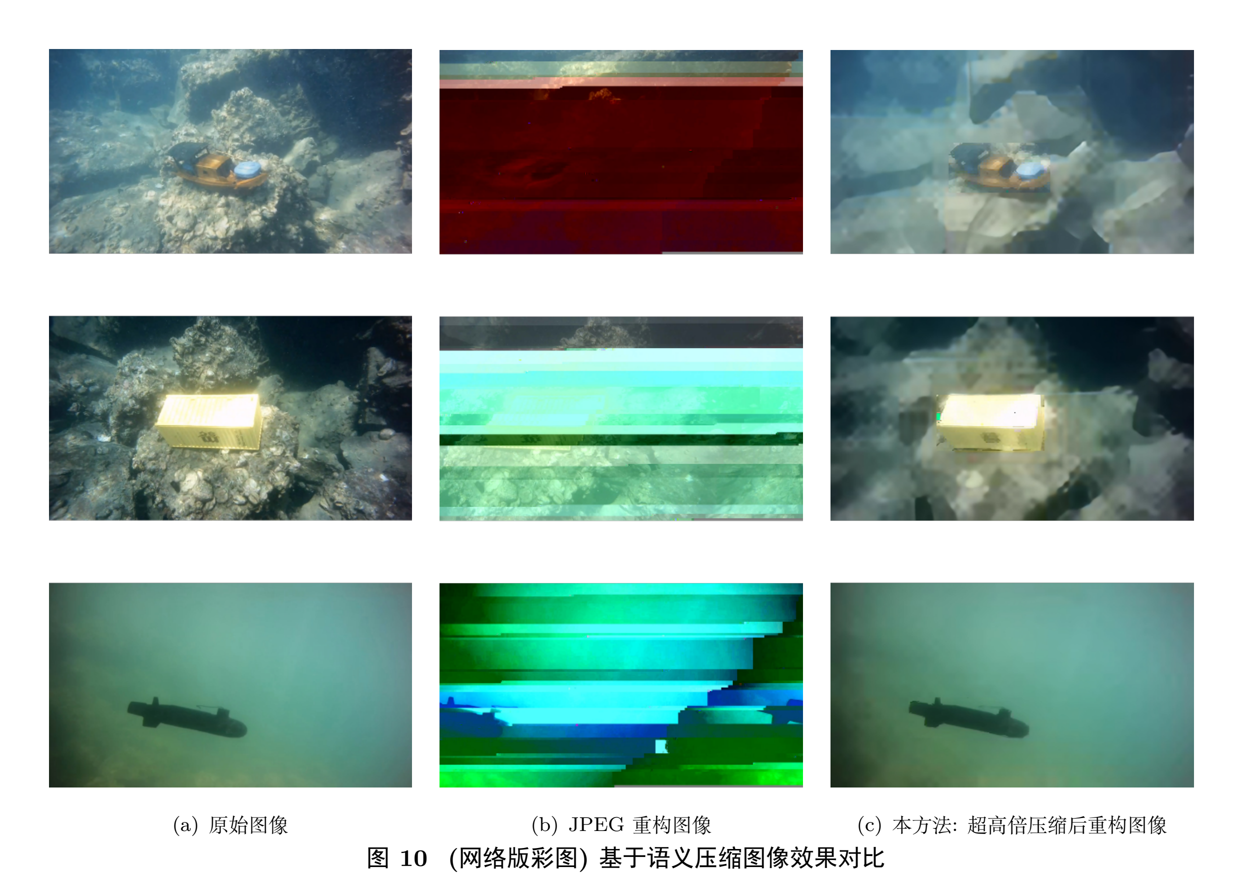
(5) 基于知识库对所有符号和其关系进一步凝练, 形成简化表达.

■ 解码算法

(1) 依托知识库, 展开简化符号表达, 形成具体表达;

(2) 依托语义的物景-像景关系以及符号的时空关系, 将符号转化成物景信号;

(3) 依托物景知识库, 渲染物景信号, 提高重构信号的展示效果.



(a) 原始图像　(b) JPEG 重构图像　(c) 本方法: 超高倍压缩后重构图像

图 10　(网络版彩图) 基于语义压缩图像效果对比

Figure 10　(Color online) Comparison of image quality based on semantic compression.

其算法过程如图 9所示.

首先将 $y(x,t)$ 物景信号、或像景信号以语义特征为先验压缩成语义符号 $[S_i]$, 再在语义知识库的支撑下, 进一步凝练语义符号形成高层次语义符号 $[S_j]$, 从而实现了超高比例信号压缩, 而且有超高的语义保真率. 其算法原理结构图如图 9(a)所示, 解码过程相反如图 9(b)所示. 其原理是因为内容可以用代表表示, 知识库是很好的代表. 在信源信号中:

$$y(x,t) = \sum_{i=1}^{n} f_i(x,t) + \sum_{j=1}^{k} D_j(x,t) + n(x,t) \tag{14}$$

有语义特征信号 $f_i(x,t)$、非语义特征信号 $D_j(x,t)$ 以及非特征信号 $n(x,t)$ 等分量信号. 只有语义特征信号 $f_i(x,t)$, 才能被接收者识别成内容, $D_j(x,t)$ 和 $D_j(x,t)$ 都不是内容. 将 $f_i(x,t)$ 转成语义符号, 并在知识库的支撑下提升为高层语义符号, 大大提高了信号压缩能力且保真了内容. 假设基于语义符号的重构信号为 $\hat{y}(x,t)$, 则 $\hat{y}(x,t)$ 与 $y(x,t)$ 在信号层面上有很大的误差, 而在内容层面上没有差别. 所以实现了 "信号失真而内容保真" , 或称为 "信号有损, 信息无损" . 当然对于非语义特征信号 $D_j(x,t)$, 采用本文第 4节中的非语义特征信号的分解方法在子语义特征库内分解, 直到找到子或子子语义特征信号, 则据此定义 $D_j(x,t)$ 的属性, 并定义新语义符号表达 $D_j(x,t)$.

针对水下相机拍摄到的目标如沉船、碍航物 (礁石、箱体) 等图像, 本文挑选了 210 张图像, 分为 9 组, 按图 9所示的方法对其压缩, 然后再重构, 结果如表 1, 在语义保真率达到 95% 条件下, 基于语义的信号压缩倍数较 JPEG 压缩倍数普遍提升了 20 倍以上. 图 10展示了三组对比效果图图, 对应的 JPEG 压缩图以及本文方法压缩的重构效果图.

9 结束语

本文从物理与数学的视角对语义信息概念进行解释和刻画, 深入探讨语义一致性度量以及语义保真率等计算方法. 在此基础上, 剖析消息、信息与知识之间的本质差异, 并针对信号的内容提取与表示提出理论性方法, 促使机器能够如同人脑一般理解信号内容并获取信息. 此外, 本文还对语义采样和语义分解提出了独特观点, 同时给出了基于语义的信号编码与解码方法, 为构建语义信息处理和传输理论奠定了坚实基础. 这篇文章是学习语义信息处理理论不可或缺的重要文献.

致谢

本文是作者本人与我的团队反复讨论的结果. 感谢他们！该项目得到国家自然基金委、重点研发计划、国家重大专项的支持.

参考文献

1 Wiener N. Cybernetics: or Control and Communication in the Animal and the Machine. Cambridge: The MIT Press, 1948
2 Shannon C E. A mathematical theory of communication. The Bell system technical journal, 1948, 27: 379-423
3 Shi G M, Li Y Y, Xie X M. Semantic communications: Outcome of the intelligence era. Int J Pattern Recogn, 2018, 31: 91-99[石光明, 李莹玉, 谢雪梅. 语义通讯　智能时代的产物. 模式识别与人工智能, 2018, 31: 91-99]
4 石光明, 杨[illegible]曦, 高大化, 等. 面向语义信息直传的通信架构. 通信学报, 2023, 44: 15-27
5 石光明, 肖泳, 李莹玉, 等. 面向万物智联的语义通信网络. 物联网学报, 2021, 5: 26-36
6 Shi G M, Xiao Y, Li Y Y, et al. From Semantic Communication to Semantic-Aware Networking: Model, Architecture, and Open Problems. Ieee Commun Mag, 2021, 59: 44-50
7 Xiao Y, Sun Z J, Shi G M, et al. Imitation learning-based implicit semantic-aware communication networks: Multi-layer representation and collaborative reasoning. Ieee J Sel Area Comm, 2022, 41: 639-658
8 Ma S, Zhang C H, Shen B, et al. Semantic feature division multiple access for multi-user digital interference networks. Ieee T Wirel Commun, 2024, 23: 15230-15244
9 Ma S, Zhang Z, Wu Y L, et al. Features disentangled semantic broadcast communication networks. Ieee T Wirel Commun, 2024, 23: 6580-6594
10 Wang Y B, Ma S, Gao D H, et al. Swin Kansformer-Based Semantic Communication Systems for Wireless Image Transmission. In: Proceedings of the 2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops). 2024, 265-270
11 Zhang P, Xu W J, Liu Y M, et al. Intellicise wireless networks from semantic communications: A survey, research issues, and challenges. Ieee Commun Surv Tut, 2024, 1-1
12 Ping Z, Xiaodong X, Chen D, et al. Intellicise communication system: model-driven semantic communications. Journal of China Universities of Posts and Telecommunications, 2022, 29: 2
13 牛凯, 戴金晟, 张平, 等. 面向 6G 的语义通信. 移动通信, 2021, 45: 85-90
14 张亦弛, 张平, 魏急波, 等. 面向智能体的语义通信: 架构与范例. 中国科学: 信息科学, 2022, 52: 907–921
15 HUBEL D H, WIESEL T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 1962, 160: 106-154
16 Huth A G, De H, Wendy A G, et al. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 2016, 532: 453-458
17 Binder J R, Desai R H. The neurobiology of semantic memory. Trends Cogn Sci, 2011, 15: 527-536
18 Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of naacL-HLT, 2019, 1: 4171-4186
19 Mikolov T. Efficient estimation of word representations in vector space. arXiv, 2013, arXiv.1301.3781
20 Patterson K, Nestor P J, Rogers T T. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci, 2007, 8: 976-987
21 Pulvermüller F. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. Trends Cogn Sci, 2013, 17: 458-470
22 Radford A K, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International conference on machine learning, 2021. 8748-8763
23 李友善, 第一性原理. 北京: 人民邮电出版社, 2021. 7
24 Vaswani A. Attention is all you need. In: Proceedings of the 31st Advances in Neural Information Processing Systems, 2017.
25 Papineni K. BLEU: a Method for Automatic Evaluation of Machine Translation, In: Proceedings of the 40th Actual Meeting of the Association for Computational Linguistics, 2002. 311–318

On Semantic Information

Guangming SHI[1,2,3*] & Dahua GAO[2,2]

1. *PengCheng Laboratory, Shenzhen* 518055*, China*;
2. *School of Artificial Intelligence, Xidian University, Xi'an* 710071*, China*;
3. *Pazhou Lab, Huangpu, Guangzhou* 510555*, China*
* Corresponding author. E-mail: gmshi@xidian.edu.cn

Abstract Claude Elwood Shannon proposed the expression of information in terms of syntax, semantics, and pragmatics as early as the 20th century. However, due to the lack of a good mathematical characterization of semantics at that time, information technology has remained focused on the syntactic level. As a result, information technology has primarily dealt with signal perception, transmission, and processing, lacking direct theoretical methods for acquiring, transmitting, and processing the content of signals. A semantic gap has always existed between the signal and its content. Signals can be represented by mathematical functions, forming the foundation of signal processing theory based on Shannon's information theory, Nyquist sampling theorem, and Fourier transform methods. However, the semantics and content of signals have not yet been well mathematically represented, making it difficult to bridge the semantic gap, let alone process semantic information. Bridging this gap has been a common research topic across the fields of computing, information, and intelligence. As society enters the era of intelligence, human-machine interaction scenarios are approaching. In particular, with the rise of semantic communication concepts and technologies, enabling intelligent machines to understand signal content has become a key issue in intelligent technology. Many scholars from engineering universities, research institutes, and developers from major enterprises have shown strong interest in semantic communication. However, from a professional perspective, the concept of semantic information remains unclear, lacking a unified and accepted definition and characterization, and there are even incorrect viewpoints. Moreover, there is no mathematical characterization of signal content. This paper discusses the connotations of information, provides clear definitions and specific calculation methods for the basic concept of semantic information, the physical generation process of semantic information, the characterization and measurement of semantic information, as well as semantic-based signal representation, compression, and the mathematical characterization of signal content. It aims to form a theory of semantic information processing to deepen and strengthen the theoretical foundation of intelligent communication and AI technology.

Keywords Semantic information, Semantic communication, Semantic characterization, Semantic compression, Signal content